17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Contextual Information based Network with High-Frequency Feature Fusion for High Frame Rate and Ultra-Low Delay Small-Scale Object Detection

Dongmei Huang[1], Jihan Zhang[1], Tingting Hu[1,2], Ryuji Fuchikami[2], Takashi Ikenaga[1]

[1]Graduate School of Information, Production and Systems, Waseda University
Kitakyushu 808-0135, Japan
[2]Panasonic Corporation
Fukuoka 812-8531, Japan
kotopai@asagi.waseda.jp

Abstract

High frame rate and ultra-low delay small-scale object detection plays an important role in factory automation for its timely and accurate reaction. Although many CNN based detection methods have been proposed to improve the accuracy of small object detection for the low resolution and large gap between the object and the background, it is difficult to achieve a trade-off between accuracy and speed. For the pursuit of ultra-low delay processing by utilizing FPGA, this paper proposes: (A) IoU and distance based loss function, (B) Contextual information with high temporal correlation based parallel detection, (C) High frequency feature fusion for enhancing low-bit networks. The proposed methods achieve 45.3 % mAP for test sequences, which is only 0.7 % mAP lower compared with the general method. Meanwhile, the size of the model has been compressed to 1.94 % of the original size and reaches a speed of 278 fps on FPGA and 15 fps on GPU.

1. Introduction

Small object detection is an important part of promoting industrial and agricultural automation. There are many applications that use small object detection, such as defective product detection in assembly line. The high frame rate means that the detector provides more detailed position information in the same time compared to others. Ultra-low delay means that the detector reacts to the deviation of the detected object in time. The combination provides more timely positioning and facilitate subsequent processing.

The definition of the small-scale object in this paper is that the area of the object in the image is less than 32*32 pixels [1]. The low resolution of the object itself leading to the few high-level semantic information after convolutions. In addition, the huge size gap between the object and the background also makes it difficult to select the receptive field. Therefore, although there are many detectors that perform well on medium and large objects, they perform poorly on small objects [2][3].

At present, in addition to some data set enhancement methods like creating sufficient positive samples [4], the accuracy of small object detection is mainly improved from three aspects: multi-scale feature map fusion, context information fusion and GAN-based methods. Multi-scale feature fusion is mainly to resize the image [5] to form an image pyramid or to extract feature maps of different layers to form a pyramidal feature hierarchy [6]. Fusion of contextual information utilizes the correlation between the detected object and the space in which it is located. For example, when detecting faces, the neck and shoulders are also included in the detection range [7]. The GAN-based method uses additional adversarial networks to enhance the features of small objects, so that their representation is similar to the feature maps of large objects [8]. Although these methods are beneficial to the improvement of accuracy, they also increase the complexity and the overall amount of calculation.

Nowadays, methods such as quantization [9] and pruning [10] are commonly used to simplify the calculation and speed up the processing, but this is not enough for realizing a high frame rate and ultra-low delay system. Zhang et al. [11] directly mapped a dense binary network to the FPGA with a hard-wired type implementation, and realized ultra-low delay dual-hand tracking. it is, however, difficult for the binary network to provide effective accuracy for the prediction of the bounding box. Therefore, the detection of small objects with high frame rate and ultra-low latency is still a challenging problem.

As the first attempt to achieve ultra-low delay processing of complex operations, this paper proposes three proposals. In order to avoid increasing the complexity of the network structure, (A) IoU and distance based loss function is proposed to improve the accuracy by adjusting the training strategy. For the problem of high frame-level delay in two-stage detection, (B) contextual information with high temporal correlation is used to increase the processing parallelism. For the problem of reduced accuracy after the entire network is compressed, (C) high frequency feature fusion for enhancing low-bit networks is proposed. More feature maps obtained from FPGA are fused as input for subsequent processing.

The rest of this paper is organized as follows. Section 2 explains the proposed methods. Section 3 provides the evaluation results and Section 4 summarizes the paper.

2. Proposals

With a relatively accurate detection rate and a computing structure that can be accelerated by parallel, this paper chooses Faster-RCNN [12] as the baseline. Figure 1 shows the concept of the framework. By processing the first stage of small objects at a high frame rate to increase the amount of information, and then processing the second stage at a normal speed in a hybrid manner, the detection performance is improved.

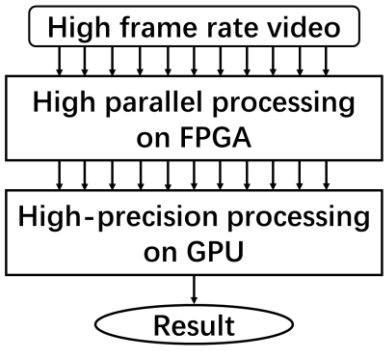

Figure 1. Concept of the framework.

2.1. IoU and Distance Based Loss Function

In object detection, the intersection over union (IoU) [13] is most commonly used as the variable of the l1-norm loss function.

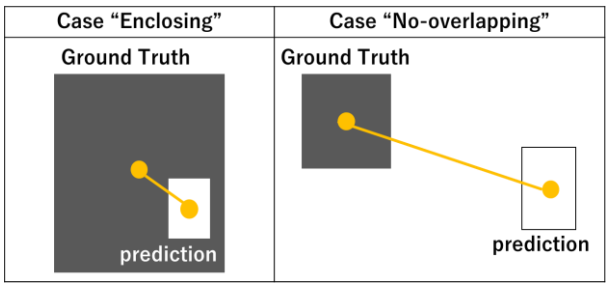

Figure 2. Two cases that often occur in small object detection

But because the ground truth and prediction of small-scale objects are small, there are often cases where they are either completely enclosed or completely no-overlapping. With only IoU, it is difficult to deal with these two situations, so the loss function based only on IoU has a worse impact on the accuracy of small object detection. Figure 2 shows two specific cases.

(1) In case Enclosing, IoU equals to the ratio between the areas of prediction and ground truth. Only the predicted size is optimized after training and the position is ignored.

(2) In case No-overlapping, IoU equals to 0 no matter how far is the prediction and ground truth.

Therefore, this paper proposes to replace the IoU with the distance of the center point after normalization in the detection part. The normalized distance is calculated by

$$dis_{norm} = sum\left(\frac{x_{gt}-x_{pred}}{W_{pred}}\right) + \text{sum}\left(\frac{y_{gt}-y_{pred}}{H_{pred}}\right), \quad (1)$$

where (x_{gt}, y_{gt}) represent the center point of the ground truth, (x_{pred}, y_{pred}) and (W_{pred}, H_{pred}) represent the center point and the size of the prediction.

By using the normalized distance, there is no need to perform repeated calculations. Both of the size and the position is updated after one training epoch

2.2. Contextual Information based Parallel Detection

In Faster-RCNN, there are three frame-level processing modules, which are the feature extraction part, the ROI generation part, and the detection part. Every time, the second frame-level processing needs to wait for the first frame-level processing, which leading to the long delay.

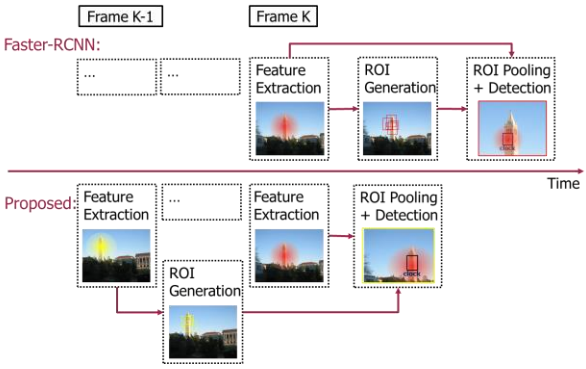

Figure 3. Concept differences between Faster-RCNN and proposed method

In order to reduce the processing delay, a high-parallel detection method based on the characteristics of high frame rate images has been proposed. Figure 3 shows the concept differences of the network structure between Faster-RCNN and proposed method.

Because the video captured by a high frame rate camera has a high temporal correlation, the second frame-level processing of current frame uses information of the first frame-level processing with little loss of accuracy. That means, when the kth frame is in the detection part, its input is the ROIs generated from the image of the k-1th frame and the feature maps extracted from the kth frame.

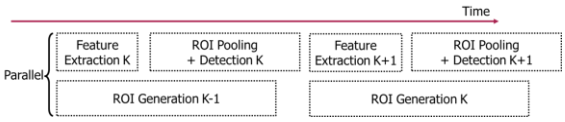

Figure 4. Parallelism of the proposed method

According to the recording of the time consumed by each part, the overall processing of the proposed method is shown in Figure 4. By parallel processing the feature extraction and detection part and the ROI generation part, the processing time will be reduced to about half of the original.

2.3. High-Frequency Feature Fusion for Enhancing Low-Bit Networks

In Faster-RCNN, the part with the largest network size is the feature extraction part. In order to further reduce the magnitude of the network size, this paper first per-

forms channel pruning and weight pruning [7] for the feature extraction part, and then quantizes it to 4 bits [6]. The parameter amount of the feature extraction part on GPU is compressed from 7.6352M to 0.1483M, which is reduced to 1.94 % of the original. While the size of the model has been greatly reduced, the accuracy has also declined from 51.1 % mAP to 41.5 % mAP.

In order to improve accuracy without increasing network complexity, this paper proposes high-frequency feature fusion based on FPGA and GPU heterogeneity. By using a smaller network in FPGA to increase temporal resolution instead of spatial resolution, which is advantageous for small object.

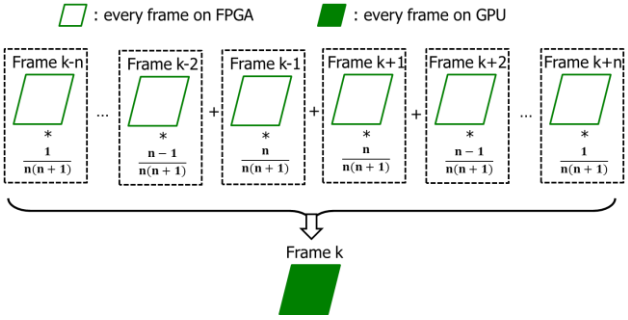

Figure 5. How the feature map obtained by FPGA replaces the feature map of GPU

When the GPU needs the feature map of the kth frame, the FPGA extracts the feature maps of n frames adjacent to the kth frame for fusion. This operation like Figure 5 shows is used to replace feature extraction on the GPU.

Because the closer to the kth frame, the higher the similarity, so each feature map extracted by FPGA needs to be given different weights when fused. The closer, the greater the weight. At the same time, when the prediction accuracy of the frame before the kth frame is less than a certain threshold, the feature map of this frame will not participate in the fusion to avoid degraded results.

The fused feature map effectively retains the image information of the kth frame itself for subsequent detection. At the same time, when there are errors and omissions in the video sequence, the feature map is more robust.

3. Experiments

3.1. Dataset and test sequences

Small object dataset [4] whose relative area of the detected object to the image is about 0.08%~0.58% is used for training. In order to avoid the problem of marking annotations, the four test sequences are made from one image randomly selected from the test set. This paper uses a rectangle to crop the part of images which have obtained the detection objects and slowly move this rectangle so that the intercepted part simulates the characteristics of a high frame rate video. Annotations are adjusted based on the number of moved pixels.

In order to deal with various possible situations, four sequences of "slowly moving", "suddenly shaking", "zoom" and "illumination change" have been produced like Figure 6 shows. Each test sequence has 1000 images.

Figure 6. Samples of test sequences in different situations

3.2. Evaluation results

The results of the evaluation are mainly divided into three aspects, that is, accuracy, speed and FPGA hardware resource consumption.

Accuracy: The impact of each proposed method on accuracy is shown in Table 1 where proposal A, B and C respectively correspond to the three proposed methods in Section 2. With proposal A, the average mAP has risen from 46.2 % to 54.4 %. When Proposal B increases the parallelism of the network structure, the accuracy is reduced to 51.1 %. The compression of the network greatly reduces the accuracy, especially for the test set of illumination changes. However, by using Proposal C with high-frequency feature fusion to strengthen the low-bit network, the accuracy has rebounded from 41.5 % to 45.3 %.

Speed: At present, the software environment is Nvidia RTX 1080Ti and the hardware environment is Xillinx Kintex-7 XC7K325T FPGA board. The original overall detection speed of faster-rcnn is 5fps. By combining the proposals, currently the feature extraction part on the FPGA has reached a processing speed of 278 fps, and the ROI generation part and the detection part on the GPU have reached a processing speed of 15 fps.

Hardware resource: The estimated hardware resource consumption of the proposed feature extractor is shown in Table 2. Although excessive hardware resources are currently consumed, and there is room for further improvement.

Table 1. Accuracy changes for each proposal

	Faster -RCNN	Proposal A	Proposal A Proposal B	Proposal A Proposal B Compression	Proposal A Proposal B Compression Proposal C
mAP (Slowly moving)	0.467	0.587	0.532	0.454	0.506
mAP (Suddenly shaking)	0.438	0.541	0.518	0.489	0.515
mAP (Zoom up)	0.476	0.519	0.494	0.431	0.487
mAP (Illumination change)	0.458	0.532	0.502	0.286	0.304
mAP (Average)	0.462	0.544	0.511	0.415	0.453

Table 2. Estimated hardware resource consumption

	Channels		Bits		Operations			FPGA resource		
	Input	Output	FIX_FM	FIX_WT	multiplier	adder	Line-buffer	LUT	DSP	BRAM
CONV-3x3	3	64	4	4	1728	1728	2	0	1728	6
DW-CONV-3x3	64	64	4	4	576	576	2	4608	0	128
PW-CONV-1x1	64	128	4	4	8192	8192	0	65536	0	0
DW-CONV-3x3	128	128	4	4	1152	1152	2	9216	0	256
PW-CONV-1x1	128	256	4	4	32768	32768	0	262144	0	0

4. Conclusion and future work

Targeting at high frame rate and ultra-low delay small-scale object detection, this paper proposes a IoU and distance based loss function. By using the high temporal correlation of high frame rate video, a high-parallel network combined with high-frequency feature map fusion is proposed to improve the detection speed while maintaining accuracy. The proposed methods achieve 45.3 % mAP for test sequences while the size of the model has been compressed to 1.94 % of the original size. The processing speed reaches 278 fps on FPGA and 15 fps on GPU.

The future work focuses on two aspects. The first one is to reduce the usage of hardware resources. The second is to find out the reason why the accuracy of the test sequence of "illumination change" drops sharply after compression and make targeted adjustments.

Acknowledgement

This work was supported by KAKENHI (21K11816).

References

[1] T.Y. Lin,M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick, P. Dollár.: “Microsoft COCO: common objects in context”, European Conference on Computer Vision 2014, pp. 740–755.

[2] M. Kisantal, Z. Wojna, J. Murawski, et al.: “Augmentation for small object detection”, arXiv:1902.07296, 2019

[3] N. Nguyen, T. Do, T. Ngo, et al.: “An Evaluation of Deep Learning Methods for Small Object Detection”, Journal of Electrical and Computer Engineering 2020, vol 2020.

[4] Y. Liu, P. Sun, N. Wergeles, Yi Shang, et al.: “A survey and performance evaluation of deep learning methods for small object detection”, Expert Systems with Applications 2021, vol 172.

[5] Singh, L.S. Davis.: “An analysis of scale invariance in object detection SNIP”, IEEE Conference on Computer Vision and Pattern Recognition 2018, pp. 3578–3587.

[6] T.-Y. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, S.J. Belongie.: “Feature pyramid networks for object detection”, IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 936–944.

[7] C. Chen., MY. Liu., O. Tuzel, J. Xiao.:“R-CNN for Small Object Detection”, Computer Vision ACCV 2016, vol 10115.

[8] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan.: “Perceptual generative adversarial networks for small object detection”, IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 1951–1959

[9] B. Jacob, S. Kligys, B. Chen, et al.: “Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2704-2713

[10]H. Li, A. Kadav, I. Durdanovic, et al.:“Pruning Filters for Efficient ConvNets”, arXiv:1608.08710, 2016

[11]P. Zhang, D. Luo, S. Du, and T. Ikenaga.:“Hetero Com-

plementary Networks with Hard-Wired Condensing Binarization for High Frame Rate and Ultra-Low Delay Dual-Hand Tracking", 13th International Conference on Human System Interaction, 2020, pp. 82-87

[12]S. Ren, K. He, R. Girshick, J. Sun.: "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", arXiv:1506.01497, 2015

[13]J. Yu1, Y. Jiang, Z. Wang, et al.: "UnitBox: An Advanced Object Detection Network", arXiv:1608.01471v, 2016